HHS Public Access

Author manuscript

Stat Med. Author manuscript; available in PMC 2016 September 10.

Published in final edited form as:

Stat Med. 2015 September 10; 34(20): 2844–2857. doi:10.1002/sim.6515.

A sexually transmitted infection screening algorithm based on semiparametric regression models

Zhuokai Li[a], **Hai Liu**[b], and **Wanzhu Tu**[b,*]

[a] Duke Clinical Research Institute, 2400 Pratt Street, Durham, NC 27705

[b] Department of Biostatistics, Indiana University Schools of Medicine and Public Health, 410 West 10th Street, Indianapolis, IN 46202

Abstract

Sexually transmitted infections (STIs) with *Chlamydia trachomatis*, *Neisseria gonorrhoeae*, and *Trichomonas vaginalis* are among the most common infectious diseases in the United States, disproportionately affecting young women. Because a significant portion of the infections present no symptoms, infection control relies primarily on disease screening. However, universal STI screening in a large population can be expensive. In this paper, we propose a semiparametric model-based screening algorithm. The model quantifies organism-specific infection risks in individual subjects, and account for the within-subject interdependence of the infection outcomes of different organisms and the serial correlations among the repeated assessments of the same organism. Bivariate thin-plate regression spline surfaces are incorporated to depict the concurrent influences of age and sexual partners on infection acquisition. Model parameters are estimated by using a penalized likelihood method. For inference, we develop a likelihood-based resampling procedure to compare the bivariate effect surfaces across outcomes. Simulation studies are conducted to evaluate the model fitting performance. A screening algorithm is developed using data collected from an epidemiological study of young women at increased risk of STIs. We present evidence that the three organisms have distinct age and partner effect patterns; for *C. trachomatis*, the partner effect is more pronounced in younger adolescents. Predictive performance of the proposed screening algorithm is assessed through a receiver operating characteristic (ROC) analysis. We show that the model-based screening algorithm has excellent accuracy in identifying individuals at increased risk, and thus can be used to assist STI screening in clinical practice.

Keywords

bivariate surfaces; multiple binary outcomes; penalized likelihood; splines; resampling

1. Introduction

Chlamydia trachomatis (CT), *Neisseria gonorrhoeae* (GC), and *Trichomonas vaginalis* (TV) are pathogenic organisms that cause sexually transmitted infections (STIs) chlamydia, gonorrhea, and trichomoniasis. Together, the three account for millions of new STI cases

* Correspondence to: Department of Biostatistics, Indiana University Schools of Medicine and Public Health, 410 West 10th Street, Indianapolis, IN 46202. wtu1@iu.edu.

each year in the United States [1, 2]. Untreated genital infections are known to have sequelae including but not limited to pelvic inflammatory disease, ectopic pregnancy, and infertility [3]. Because the infections tend to be asymptomatic (especially in women), they can linger for months or years without being detected, causing irreversible damage to the reproductive organs, and ultimately leading to infertility. From a public health standpoint, infected individuals unbeknownst of their infection status do not receive timely treatment and continue to pass the pathogens to their partners, causing the disease to multiply in a population. Therefore, an important focus of the STI prevention strategy has been the promotion of infection screening. Indeed, Centers for Disease Control and Prevention (CDC) and the US Preventive Service Task Force (USPSTF) recommend that all sexually active women under 25 years of age be screened annually for CT, and individuals with known risk factors (such as multiple partners, an infection history, etc.) be screened for GC [4, 5]. While DNA-based nucleic acid amplification tests (NAATs) are highly sensitive [6, 7], universal screening for all sexually active women in a large population can be cost inhibitive. In resource limited settings where STIs are more prevalent, such as inner city communities, the screening recommendations are not always implemented. To alleviate the burden of universal screening, attempts have been made to focus on women with specific risk factors, especially those with prior infections or multiple partners. However, these efforts have only resulted in limited successes because of suboptimal sensitivity and specificity [8, 9].

In this paper, we propose a model-based screening algorithm to target individuals at higher risk. The model incorporates the effects of known risk factors into the estimation of organism-specific STI probabilities. Here, we construct the model with the following considerations: (1) The ability to account for the synergistic relationships among the three organisms. It has been well documented that co-infections with multiple organisms are common, especially between CT and GC. In certain populations, up to 70% of GC-positive youths were co-infected with CT [10, 11]. Our own data in adolescent women suggested that CT, GC and TV infections tended to cluster, possibly due to the organisms' biological synergy and their common mode of transmission [12, 13]; this provides a compelling rationale to consider a joint modeling approach. (2) Accommodation of possible nonlinear effects of risk factors on STI acquisition. Previous studies suggested that younger adolescents were at greater risk for STI, particular with CT [14]. Nonetheless it is unclear whether a linear age effect is adequate to quantify an individual's STI risk. Similarly, having multiple sexual partners is a strong predictor for STI acquisition [15, 16], but our own data suggested STI risk did not increase linearly with the number of partners, possibly due to the increased prophylactic use in individuals with multiple partners [17]. (3) Accommodation of potentially interacting influences. Effects of STI risk factors are unlikely to be additive. For example, a woman's infection risk depends not only on the behaviors that expose her to a source of infection, but also on her own immunological response to the disease pathogen [18]. If the number of partners marks the level of exposure, strength of host immune response may be related more to the host biological conditions such as age. It is therefore important that a targeting algorithm correctly depicts these interacting influences. Aggregating the aforementioned features into a statistical model, we envision a joint semiparametric logistic regression model with bivariate independent variable effects.

Herein, the joint modeling structure is used to connect organism-specific infection outcomes; the semiparametric bivariate effects are used to accommodate the nonlinear and potentially interacting influences. We contend that such a targeting algorithm may have the potential to significantly improve screening efficiency.

Methodologically, constructing and fitting such a model is not trivial. To the best of our knowledge, no existing models have all of the required features. This said, various components of the model have been developed in other contexts. For example, two general approaches in multivariate regression analysis of longitudinal data have been developed. One is based on the generalized estimation equation (GEE) techniques [19, 20], including applications to binary outcomes [21, 22]. The other is the mixed effects model [23, 24]. Various semiparametric models have also been proposed for nonlinear independent variable effects on multiple outcomes. Coull and Staudenmayer [25] described a self-modeling regression method which extended penalized regression splines [26] to multivariate longitudinal data. Ghosh and Tu [27] proposed a joint semiparametric model for zero-inflated counts that consisted of a logistic model for the proportion of zeros and a log-linear model for Poisson counts, and both models included univariate nonparametric components. Ghosh and Hanson [28] developed a semiparametric Bayesian approach for multivariate longitudinal data, in which the conventional normal assumption for random effects was relaxed. More recently, Liu and Tu [29] developed a joint semiparametric model for paired continuous outcomes, which incorporated bivariate smooth components.

In this paper, we propose a joint semiparametric regression model for longitudinal binary data, to model infection acquisition of different organisms. The model accounts for correlations across the organisms and that among the repeated measurements of the same organism over time. Joint modeling of multiple outcomes is accomplished by specifying a covariance structure for the shared random effects. Additionally, bivariate smoothing components are incorporated into the model for nonlinear effects and their interactions. We use the model to quantify the organism-specific infection risks, and assess the predicative accuracy of the model using a receiver operating characteristics (ROC) analysis.

2. Data Source

To develop the STI screening algorithm, we use data from a longitudinal cohort of inner-city young women, hereafter referred to as the Young Women's Project (YWP). The study was approved by a local Institutional Review Board and its protocol was described elsewhere [30]. Briefly, young women aged 14 to 17 attending three primary care clinics were recruited for participation in this observational study. At enrollment, the participants were tested for CT, GC, and TV infections; those infected were treated promptly. They also completed an interview on their lifetime and most recent sexual behaviors, including the number of sexual intercourse, condom use, and the number of sexual partners in the last three months. The participants returned to clinic every three months, at which time they had face-to-face interviews and received STI tests. Infections identified at all follow-up visits were considered as incident cases (i.e., newly acquired infections) because all prior infections were treated. The mean length of follow-up was approximately 3.2 years; the longest follow-up was 7.8 years. Of 5,213 follow-up visits of all participants, CT, GC and

TV infection status were missing at only 20, 23 and 1 visit(s), respectively. A high completion rate for quarterly interviews was also achieved, with only 5% of possible follow-up interviews missing.

The study sample included 386 young women, consisting of 344 (89.1%) African Americans, 39 (10.1%) Whites and 3 (0.8%) Hispanics. Co-infections with different organisms were common in the study sample. Of 193 cases of GC infection, 31.6% were co-infected with CT, and 14.5% were co-infected with TV; of 287 cases of TV infection, 16.0% were co-infected with CT. At enrollment, the participants were between 14 and 17 years of age, with a mean age of 15.8 years and a standard deviation of 1.1 years.

We examined the relationship between age and the number of sexual partners in the study participants, and found that the number of partners increased with age in early and mid-adolescence until it peaked between 19 and 20 years of age. Figure 1 shows the infection rates of CT, GC and TV by age group and the number of sexual partners in the last 3 months. Both age and the number of partners appear to have a nonlinear relationship with all three types of infections. The age patterns across the organisms are different, with the highest infection rates occurring at ages 16 – 17, 18 – 19 and 24 – 25 for CT, GC and TV, respectively. These nonlinear patterns point to the need of nonparametric regression models. Further, by introducing bivariate smooth functions into the analysis, we hope to capture the potential interactions between age and the number of partners, which are not available for assessment in additive models.

3. Model Specification

3.1. A Joint Semiparametric Model for Binary Outcomes

Let Y_{ij}^k be the *i*th individual's infection status with sexually transmitted organism *k* at the *j*th visit, $i = 1, 2, \ldots, m$, $j = 1, 2, \ldots, n_i$, and $k = 1, 2, \ldots, K$, where m is the number of individuals, n_i is the number of follow-up visits for the *i*th individual, and K is the number of sexually transmitted organisms in the study. The infection status Y_{ij}^k is a binary outcome with $Y_{ij}^k = 1$ and $Y_{ij}^k = 0$ indicating positive and negative test results, respectively, for organism *k*.

Assuming Y_{ij}^k follows a Bernoulli distribution with parameter p_{ij}^k, we propose the following model

$$g\left(p_{ij}^k\right) = S_i^T\beta_1^k + T_{ij}^T\beta_2^k + \sum_{q=1}^{Q}\beta_{3q}^k Y_{i,j-q} + Z_{ij}^T b_i^k + f^k\left(u_{ij}, v_{ij}\right), \quad (1)$$

for $k = 1, \ldots, K$, where $g(\cdot)$ is a known invertible link function, e.g., logit link. The parameter vectors β_1^k and β_2^k represent respectively the fixed effects regression coefficients associated with time-independent covariates $\boldsymbol{S}_i$ and time-dependent covariates $\boldsymbol{T}_{ij}$. The *q*th order autoregressive component $Y_{i,j-q}$ indicates the prior infection status with any of the *k* organisms at the $(j - q)$th visit. Let $\beta_3^k = \left(\beta_{31}^k, \ldots, \beta_{3Q}^k\right)$ denote the coefficient vector for

the autoregressive component. When the follow-up visits are approximately regularly spaced without missing data, for example, if $Q = 1$, a fixed parameter β_{31}^k is sufficient to characterize the effect of lag-1 infection on the current status. If some of the follow-up visits are irregularly spaced or missing, a time-varying coefficient $\beta_{3q}^k(t_{i,j} - t_{i,j-q})$ can be used, with $t_{i,j}$ and $t_{i,j-q}$ being the time at the jth and $(j - q)$th visits, respectively. The time-varying autoregressive structure is adopted in the analysis of the YWP data in Section 6. We also incorporate a bivariate function $f^k(u_{ij}, v_{ij})$ in order to capture the nonlinear effects of other risk factors, such as age and the number of sexual partner, and their potential interaction effects on STIs. To accommodate the interdependence of multiple organisms within an individual as well as the within-subject correlations among the repeated measurements, we introduce the random effects b_i^k into the model, which in general, can be a random vector with multivariate normal distribution. For simplicity, we assume a simple, scalar random effects term b_i^k in the context of our example. We denote the vector of subject-specific random effects by $\boldsymbol{b}_i = \left(b_i^1, \ldots, b_i^k\right)^T$, and assume that it follows a multivariate normal distribution, i.e., $\boldsymbol{b}_i \sim N_K(\mathbf{0}, \boldsymbol{\Omega}_b)$, with variance-covariance matrix $\boldsymbol{\Omega}_b$.

For each bivariate smooth function in the proposed model, we specify a set of basis functions $h_l^k = 1, \ldots, M_k$, so it can be expressed as $f^k(u, v) = \Sigma_{l=1}^{M_k} \gamma_l^k h_l^k(u, v)$, and $\gamma_k = \left(\gamma_1^k, \cdots, \gamma_{Mk}^k\right)$ denotes the vector of regression coefficients for f^k. Let $\boldsymbol{f}^k$ be a vector of smooth functions with elements $f^k(u_{ij}, v_{ij})$, for $j = 1, \ldots, n_i$; $i = 1, \ldots, m$, i.e., $\boldsymbol{f}^k = [f^k(u_{ij}, v_{ij})]_{1 \le j \le n_i;\, j \le m_i}$, then it can be written in a matrix form $\boldsymbol{f}^k = \mathbf{X}_k \gamma_k$, where the design matrix $\mathbf{X}_k = \left[h_1^k(u_{ij}, v_{ij}), \cdots, h_{M_k}^k(u_{ij}, v_{ij})\right]_{1 \le j \le n_i; 1 \le i \le m}$.

In this research, thin plate regression splines are used to model the smooth functions, which provide good approximation to the full rank thin plate splines and significantly reduce the computational cost. Furthermore, truncated eigen- decomposition is used to avoid choosing knot locations for thin plate regression splines [31, 32]. The smooth function estimators can be obtained by maximizing the penalized log-likelihood function of model (1),

$$\ell_p = \ell - \sum_{k=1}^{K} \lambda_k J\left(f^k\right) \quad (2)$$

where ℓ is the log-likelihood function of the model, and λ_k is the smoothing parameter associated with f^k, which balances goodness-of-fit and smoothness of the model. In the case of bivariate smoothing, we define the roughness penalty $J(f)$ as

$$J(f) = \iint_{\mathbb{R}^2} \left\{ \left(\frac{\partial^2 f}{\partial u_2}\right)^2 + 2\left(\frac{\partial^2 f}{\partial u \partial v}\right)^2 + \left(\frac{\partial^2 f}{\partial v}^2\right)^2 \right\} \mathrm{d}u\mathrm{d}v,$$

which can be expressed as a quadratic form in regression coefficients γ_k. For example, $J\left(f^k\right) = \gamma_k^T M_k \gamma_k/2$, where $\mathbf{M}_k$ are positive semi-definite matrices of known coefficients. Therefore, the penalized log-likelihood function (2) can be rewritten as

$$\ell = \ell - \frac{1}{2}\sum_{k=1}^{K}\gamma_k^T \Lambda_k \gamma_k, \quad (3)$$

where the penalty matrix $\mathbf{\Lambda}_k = \lambda_k \mathbf{M}_k$.

3.2. Mixed Model Representation

Semiparametric models using penalized splines can be presented as mixed effects models [26, 32], and as a result, mixed model methodology and software can be adopted for the fitting of model (1). First, the penalized smooth functions, $\boldsymbol{f}^k$, are divided into fixed and random components of a mixed effects model, which is achieved by using the eigen-decomposition of $\mathbf{\Lambda}_k$ [32]. The regression coefficient vector of $\boldsymbol{f}^k$ is written as $\gamma_k = \left(\gamma_{k,F}^T, \gamma_{k,R}^T\right)^T$, where $\gamma_{k,F}$ represents the vector of unpenalized coefficients as fixed effects, and $\gamma_{k,R}$ represents the penalized coefficients, which are regarded as random effects. The penalty matrix corresponding to $\gamma_{k,R}$ is denoted by $\mathbf{\Lambda}_{k,R}$ such that $\gamma_k^T \mathbf{\Lambda}_k \gamma_k = \gamma_{k,R}^T \Lambda_{k,R} \gamma_{k,R}$. Accordingly, the design matrix of the smooth term $\boldsymbol{f}^k$ is partitioned into two parts, $\mathbf{X}_k = (\mathbf{X}_{k,F}, \mathbf{X}_{k,R})$.

We then rewrite model (1) in the form of a generalized linear mixed model (GLMM). Let $\boldsymbol{Y} = \left(\boldsymbol{Y}_1^T, \cdots, \boldsymbol{Y}_k^T\right)^T$ be the response vector, where $\boldsymbol{Y}_k = \left[Y_{ij}^k\right]_{1\le j\le n_i; 1\le i\le m}$. The corresponding mean vector $\boldsymbol{p}$ is related to the linear predictor through a vector-valued link function $\boldsymbol{g}$. Defining $\tilde{\beta}_k = \left(\left(\beta_1^k\right)^T, \left(\beta_2^k\right)^T, \left(\beta_3^k\right)^T\right)^T$ and $\tilde{\beta} = \left(\tilde{\beta}_1^T, \ldots, \tilde{\beta}_K^T\right)^T$, we write the vector of fixed effects parameters as $\beta = \left(\tilde{\beta}^T, \gamma_{1,F}^T, \cdots, \gamma_{K,F}^T\right)^T$. Similarly, we define $\tilde{b}_k = \left(b_1^k, \cdots, b_m^k\right)^T$ and $\tilde{b} = \left(\tilde{b}_1^T, \cdots, \tilde{b}_K^T\right)^T$, and thus the vector of random effect parameters becomes $b = \left(\tilde{b}^T, \gamma_{1,R}^T, \cdots, \gamma_{K,R}^T\right)^T$. The design matrix associated with $\tilde{b}$ can be written as $\tilde{\boldsymbol{Z}} = I_K \otimes \boldsymbol{Z}_b$ such that the components of $\boldsymbol{Z}_b \tilde{b}_k$ corresponding to subject i are equal to b_i^k. The design matrix associated with $\tilde{\beta}$ is set up as follows:

$\tilde{\mathbf{X}} = \mathbf{I}_k \otimes \mathbf{X}_\beta$, where $\mathbf{X}_\beta = (\mathbf{S}, \mathbf{T}, \mathbf{Y}_Q)$, and $\mathbf{S} = \boldsymbol{Z}_b\left[\mathbf{S}_i^T\right]_{1\le i\le m}$, $\mathbf{T} = \left[\mathbf{T}_{ij}^T\right]_{1\le j\le n_i; 1\le i\le m}$ and $\boldsymbol{Y}_Q = \left[\left(\boldsymbol{Y}_{ij,Q}^k\right)^T\right]_{1\le j\le n_i; 1\le i\le m}$. Model (1) can therefore be expressed as a GLMM

$$g(p) = \mathbf{X}\beta + \boldsymbol{Z}b, \quad (4)$$

where $\mathbf{X} = \left(\tilde{\mathbf{X}}, \text{diag}\left(\mathbf{X}_{1,F}, \cdots, X_{K,F}\right)\right)$ and $\boldsymbol{Z} = \left(\tilde{\boldsymbol{Z}}, \text{diag}\left(\mathbf{X}_{1,R}, \cdots, X_{K,R}\right)\right)$ are the design matrices associated with the fixed effects and the random effects, respectively. The random effects vector $\boldsymbol{b} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_b(\boldsymbol{\theta}))$, where $\Sigma_b(\theta) = \text{diag}\left(\Omega_b \otimes \mathbf{I}_m, \Lambda_{1,R}^{-1}, \cdots, \Lambda_{K,R}^{-1}\right)$ with $\boldsymbol{\theta}$ being a vector of the variance components.

3.3. Estimation Procedure

The likelihood function of the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ up to a multiplicative constant is

$$L(\beta, \theta | y) = |\Sigma_b(\theta)|^{-1/2} \int \left[\prod_{k=1}^{k} \prod_{i=1}^{m} \prod_{j=Q+1}^{n_i} \left(p_{ij}^k\right)^{y_{ij}^k} \left(1 - p_{ij}^k\right)^{1-y_{ij}^k} \right] exp\left(-\tfrac{1}{2} b^T \Sigma_b^{-1}(\theta)\, b\right) \mathrm{d}b, \quad (5)$$

where p_{ij}^k is a function of $\boldsymbol{\beta}$ and $\boldsymbol{b}$, as defined in model (1). The integration in the likelihood function is tractable for linear mixed models where the response variable is normally distributed, but for binary outcomes it does not have a closed-form expression. Instead it can be evaluated using a Laplace approximation [33]. Note that the integrand in equation (5) is the unnormalized conditional density of the random effects $\boldsymbol{b}$ given $\boldsymbol{Y} = \boldsymbol{y}$. For given $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, the conditional mode of $\boldsymbol{b}$ is

$$\hat{b}(\beta, \theta) = \arg \max_b \left\{ \sum_{k=1}^{K} \sum_{i=1}^{m} \sum_{j=Q+1}^{n_i} \left[y_{ij}^k log\left(p_{ij}^k\right) + \left(1 - y_{ij}^k\right) log\left(1 - p_{ij}^k\right) \right] - \frac{1}{2} b^T \Sigma_b^{-1}(\theta)\, b \right\},$$

which can be determined using a penalized iteratively reweighted least squares (PIRLS) algorithm [34]. At the conditional mode $\hat{b}(\beta, \theta)$, the Laplace approximation to the likelihood $L(\boldsymbol{\beta}, \boldsymbol{\theta}|\boldsymbol{y})$ can be optimized to obtain the approximate maximum likelihood estimates (MLEs) for $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ [35]. The smoothing parameters λ_k, $k = 1, \cdots, K$, are selected implicitly through the Laplace approximated maximum likelihood.

The estimation procedure for model (1) is implemented by using package gamm4 (generalized additive mixed models using mgcv and lme4) in R [36]. An alternative estimation method for GLMMs is penalized quasi-likelihood (PQL) [35, 37] in which the likelihood is replaced by a quasi-likelihood and maximized as in a linear mixed model to obtain the approximate MLEs. For binary outcomes, however, estimates of the fixed effects and variance components of the random effects resulting from PQL tend to be biased toward zero [38–40]. Therefore, the Laplace approximation method is used in our simulation study and data analysis, for the more robust numerical performance.

4. Statistical Inference

Although prediction is the intended utility of the model, the proposed modeling structure does offer a capacity for statistical inference. For example, a question that one may be interested in is whether age and the number of partners have different effects on infection acquisition of different organisms. In model (1), the concurrent influences of the two

independent variables on the infection of organism *k* are represented by bivariate function f^k. Therefore, the question can be formulated into a hypothesis about the functional forms of f^k,

$$H_0 \ : \ f^1 \ = \ \cdots \ = f^K \ \text{ vs. } \ H_1 \ : \ \text{otherwise.} \quad (6)$$

Zhang and Lin [41] considered testing the equivalence of two nonparametric univariate functions in semiparametric additive mixed models for two groups. They constructed a test statistic based on the integrated squared difference of two functions and they approximated the distribution of the test statistic by a scaled chi-square distribution. However, it is difficult to apply the test they developed to compare bivariate smooth functions. Herein, we propose a likelihood ratio test (LRT) based on the test statistic $\Delta \ = \ -2\left[\ell\left(\hat{\beta}_0, \hat{\theta}_0\right) - \ell\left(\hat{\beta}, \hat{\theta}\right)\right]$, where $\ell\left(\hat{\beta}_0, \hat{\theta}_0\right)$ is the maximized value of the log- likelihood under H_0, and $\ell\left(\hat{\beta}, \hat{\theta}\right)$ is the maximized log-likelihood for the unrestricted model. Theoretically, it is difficult to derive the asymptotic distribution of the test statistic Δ under the null hypothesis. The asymptotic properties of LRT based on the large sample chi-squared mixture approximations are rarely satisfactory when applied to penalized splines models [42]. Therefore, we resort to resampling techniques to approximate the sampling distribution of Δ. Härdle *et al.* [43] have shown that bootstrap can be applied to componentwise hypothesis testing in semiparametric generalized additive models. Roca-Pardiñas *et al.* [44] also used a bootstrap method to test factor-by-surface interactions in a logistic generalized additive model. Liu and Tu [29] extended the bootstrap test for comparing the bivariate surfaces among different groups of subjects to a longitudinal data setting with paired outcomes. Here we use the same strategy to compare the bivariate effects across the outcomes.

To test the hypothesis in (6), we consider a resampling procedure which combines bootstrap and permutation techniques:

1. Fit model (1) under the null hypothesis to obtain the effective degrees of freedom (EDF) for the penalized splines estimates.
2. Draw a bootstrap sample with replacement from the observed data. The sampling units are individuals, that is, either none or all of the observations from an individual are selected. If an individual is selected more than once, he/she will be treated as a different person in the bootstrap data.
3. Permute the labels indicating the 1st, 2nd, … and *K*th outcomes within each individual in the bootstrap sample, preserving the order of the repeated measurements for each outcome.
4. For the bootstrap data with permuted labels, refit the null and unrestricted models using regression splines with the degrees of freedom (DF) fixed at the EDF estimated in step 1, and calculate the likelihood ratio test statistic Δ^*.
5. Repeat steps 2 − 4 for *B* times to generate a sample of test statistic $\{\Delta_b^*\}_{1 \le b \le B}$, representing an empirical distribution of Δ under the null hypothesis. The p-value is calculated as $p \ = \ \#_{b=1}^{B}\{\Delta_b^* \ge \Delta\}/B$.

In the absence of asymptotic results, the resampling procedure provides a valid alternative to the traditional large sample theory based inferences. The performance of the procedure was assessed in a simulation study described in Section 5.

5. Simulation Study

A simulation study was conducted to evaluate the performance of the proposed model fitting procedure. For the outcomes, we generated two correlated binary variables $Y_{ij}^k|b_i^k, Y_{i,j-1}^k \sim \text{Bernoulli}\left(p_{ij}^k\right)$ for $i = 1, \ldots, m$; $j = 1, \ldots, n$; $k = 1, 2$ using the following model

$$\text{logit}\left(p_{ij}^k\right) = \beta_0^k + \beta_1^k Y_{i,j-1}^k + b_i^k + \bar{f}_k\left(u_{ij}, v_{ij}\right), \quad (7)$$

where $\left(b_i^1, b_i^2\right)^T \sim \mathbf{N}\left(0, \mathbf{\Omega}_b\right)$ with

$$\Omega_b = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}.$$

In model (7), the autoregressive term $Y_{i,0}^k$ was generated from Bernoulli(0.5), u_{ij} from Uniform(0, 30), and v_{ij} from Uniform$\{0, 1, \cdots, 10\}$. We considered two different nonlinear bivariate functions $f_1(u, v) = \exp[-(u - 5)^2/200 - (v - 10)^2/50 + (u + 8)(v - 10)/300]$ and $f_2(u, v) = 0.4 \exp[-(u - 18)^2/500 - (v - 10)^2/40 + (u - 15)(v - 5)/200]$. The joint effects of (u_{ij}, v_{ij}) on the response variables had functional forms of $\bar{f}_1$ and $\bar{f}_2$, corresponding to the centered functions f_1 and f_2 over the simulated covariates, respectively. The fixed effects parameters were chosen as: $\beta_0^1 = -2.5$, $\beta_0^2 = -3.5$, $\beta_1^1 = 1$, and $\beta_1^2 = 0.7$. The parameters in the variance components were set to $\sigma_1 = 0.6$, $\sigma_2 = 1$, and $\rho = 0.7$.

We fitted model (7) to the simulated data. The model performance was then assessed under the following sample size settings: $m = 200, 400$, and $n = 10, 20$. Point estimates for the fixed effects and the variance components were averaged over 200 simulation runs. The standard errors and coverage probabilities of the 95% confidence intervals (CIs) were calculated using a bootstrap method [45]. The mean squared errors (MSEs) of the smooth function estimates $\hat{f}_1$ and $\hat{f}_2$ (subject to a centering constraint) were reported for each of the simulation settings. The R code for this simulation study and a simulated dataset are provided in Section A of the Supplement.

The simulation results are presented in Table 1. In summary, the estimation procedure performed well, and the parameter estimates approached the true values as the sample size (either the number of subjects or the number of repeated outcome measurements) increased. We note that the estimation bias in the autoregressive coefficients was significantly reduced when the number of repeated measurements increased. The coverage probabilities of the bootstrap CIs were close to the nominal level 95%. For the fixed effects, we compared the bootstrap CIs with the Wald-type CIs (not shown) constructed using the parameter estimates and standard errors reported from gamm4, and found that the bootstrap CIs provided

improved coverage probabilities for most of the parameters. The MSEs of both smooth functions steadily decreased as the sample size increased. In conclusion, the proposed model achieved a satisfactory performance in the estimation of parameters and bivariate smooth functions.

We performed an additional simulation study with three correlated binary outcomes to further evaluate the model performance. Two hundred data sets were generated using a model similar to (7) with $m = 200$ and $n = 10$. Different bivariate nonlinear functions were specified for the three outcomes. We fitted both the joint model and individual models (i.e., one model for each of the three outcomes) to the simulated data, and compared the model fitting results (Table 1 in the Supplement). The joint model resulted in reduced bias in the parameter estimates and better coverage probabilities of the 95% bootstrap CIs. The bootstrap standard errors of the parameters and the MSEs of the smooth functions estimated from the joint model were consistently smaller. The efficiency improvement was more evident for the variance components. As expected, the joint model outperformed the individual models in terms of estimation efficiency and accuracy, as the individual models ignored the interdependence among the outcomes.

Another simulation study was conducted to assess the performance of the likelihood-based resampling procedure proposed in Section 4. We generated data using the same model as in the three-outcome simulation study described above, except that the bivariate functions were assumed to have the same functional form for all outcomes. The size of the test was assessed based on 200 simulation runs, each including 200 bootstrap samples. Under a sample size of 200 subjects with 10 repeated measurements on each outcome per subject, the resampling test achieved a size of 0.04, which was close to the nominal level 0.05.

To assess the predictive accuracy of the model, we performed an ROC analysis with a 10-fold cross validation in a simulation study. The details are given in Section C of the Supplement. Simulation results indicate that the proposed semiparametric model has a much improved prediction accuracy compared to the traditional generalized linear mixed models, in the presence of nonlinear effects.

6. Data Analysis

6.1. Model Development

The YWP data described in Section 2 are used to construct the proposed model, which quantifies the organism-specific infection probabilities based on the risk factors including age, the number of sexual partners and an infection history. The data include 386 participants with a total of 5,213 follow-up visits. Let Y_{ij}^{ct}, Y_{ij}^{gc} and Y_{ij}^{tv} be the *i*th participant's infection status corresponding to CT, GC and TV at the *j*th visit, $i = 1, \ldots, 386$, $j = 1, \ldots, n_i$, and n_i ranges from 1 to 30, with a median of 13 follow-up visits per participant.

We consider the following model

$$\begin{cases} \text{logit}\left(p_{ij}^{ct)}\right) &= \beta_0^{ct}+\beta_1^{ct}\left(t_{i,j}-t_{i,j-1}\right)Y_{i,j-1}+\beta_2^{ct}x_{ij}+b_i^{ct}+f^{ct}\left(u_{ij},v_{ij}\right) \\ \text{logit}\left(p_{ij}^{gc)}\right) &= \beta_0^{gc}+\beta_1^{gc}\left(t_{i,j}-t_{i,j-1}\right)Y_{i,j-1}+\beta_2^{gc}x_{ij}+b_i^{gc}+f^{gc}\left(u_{ij},v_{ij}\right) \\ \text{logit}\left(p_{ij}^{tv)}\right) &= \beta_0^{tv}+\beta_1^{tv}\left(t_{i,j}-t_{i,j-1}\right)Y_{i,j-1}+\beta_2^{tv}x_{ij}+b_i^{tv}+f^{tv}\left(u_{ij},v_{ij}\right),, \end{cases} \quad (8)$$

and the subject-specific random effects $b_i = \left(b_i^{ct}, b_i^{gc}, b_i^{tv}\right)^T \mathbf{N}\left(0, \mathbf{\Omega}_b\right)$ where

$$\Omega_b = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \rho_{13}\sigma_1\sigma_3 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & \rho_{23}\sigma_2\sigma_3 \\ \rho_{13}\sigma_1\sigma_3 & \rho_{23}\sigma_2\sigma_3 & \sigma_3^2 \end{pmatrix}.$$

In model (8), p_{ij}^{ct}, p_{ij}^{gc} and p_{ij}^{tv} are the corresponding means of the binary response variables conditional on the random effects b_i^{ct}, b_i^{gc}, and b_i^{tv} , respectively. The predicted values of the STI probabilities can be used to determine whether an individual should be screened, and if so, for what organisms. Organism-specific intercepts are denoted by β_0^{ct}, β_0^{gc} and β_0^{tv}. Similarly, β_1^{ct}, β_1^{gc} and β_1^{tv} are the organism-specific time-varying coefficients for the first-order autoregressive component $Y_{i,j-1}$, with $t_{i,j} - t_{i,j-1}$ being the lag time between the $(j-1)$th and jth visits. We only incorporated the first-order autoregressive terms because recurrent STIs are usually regarded as a Markov process, where the current infection status depends only on the infection status at the previous visit [18]. To control for the effect of condom use, we included a binary covariate (x_{ij}) indicating whether or not the individual had unprotected sex in the last three months, with β_2^{ct}, β_2^{gc} and β_2^{tv} as the regression coefficients. Other demographic and behavioral information can also be included as needed, for example, socioeconomic status, race and age at first sexual intercourse. Bivariate functions f^{ct}, f^{gc}, and f^{tv} represent the joint effects of age (u_{ij}) and the number of sexual partners in the last 3 months (v_{ij}) on CT, GC and TV, respectively.

We fitted model (8) to the YWP data, and obtained the parameter estimates for the fixed effects and the variance components. The standard errors and the 95% confidence intervals were computed based on 200 bootstrap samples. The estimated joint effects of age and the number of partners were depicted using colored contour plots. The bivariate effect surfaces for the three organisms were compared by using the proposed resampling procedure.

6.2. Analytical Results

The model fitting results are presented in Table 2. The estimated correlation coefficients of the random effects provide a simple quantification of the strength of interdependency among binary outcomes. Our data indicate strong pairwise correlations in the random effects among different organisms, suggesting these STIs are highly correlated, especially CT and GC ($\hat{\rho}_{12}$ = 0.75, 95% CI = [0.45, 1.00]). In other words, young women at high risk for infection with one organism are very likely to be infected with other organisms. Such relationships among different organisms would not be captured if they were modeled individually, thus demonstrating the usefulness of the proposed joint modeling approach. After examining the lag time effects of a prior infection of any type on the current infection status (Figure 1 in

the Supplement), we learn that a prior infection significantly increases the risks of CT and TV infections.

In Figure 3, the estimated bivariate surfaces of age and the number of partners are plotted with (right panel) or without (left panel) a prior infection of any type at the previous visit. The surfaces have very different shapes across the organisms. To formally test such differences, we repeated the resampling procedure 500 times. The test statistic $\Delta = 59.6$ with a p-value < 0.001, indicating a highly significant difference in the joint effects of age and the number of partners across the three organisms.

Several observations can be made from the contour plots. First, age effect has a nonlinear pattern for CT and GC. CT infection risk peaked at younger ages between 14 and 16, and then decreased steadily after age 18. GC infection risk increased until age 19, and then gradually decreased. In contrast, TV infection risk increased almost linearly with age. Second, the number of partners is a highly significant risk factor for all of the three organisms, though its effect tends to depend on the age of the individual. Specifically, having multiple sexual partners had a stronger effect on CT infection at younger ages, which means younger girls having multiple partners were more vulnerable to CT infection than older ones with the same number of partners.

6.3. Prognostic Accuracy Assessment

We performed an ROC analysis to assess the prognostic accuracy of model (8). The probability of organism-specific infections was predicted for each participant at each visit using the model constructed above. Comparing to the observed infection status, we calculated the sensitivity and specificity of the model under different cutoff points of infection probabilities, and plotted an ROC curve for each type of infection.

The ROC curves are shown in Figure 2. The areas under the curve (AUC) for CT, GC and TV are respectively 0.80, 0.87 and 0.89, indicating that the proposed model achieved excellent prognostic accuracy. To further assess the predictive accuracy, we also employed a 10-fold cross validation and obtained respective AUC of 0.77, 0.80 and 0.86 for the three organisms. As a targeted screening tool, the proposed model was able to correctly identify most individuals at high risk for further STI testing. Table 3 provides the sensitivity and specificity of the model-based screening algorithm under different cutoff points for the three organisms, and the corresponding percentages of follow-up visits that meet those cutoff points. In general, one hopes to have a highly sensitive screening algorithm to target high-risk individuals for formal STI testing while letting the low specificity be compensated by the diagnostic test. Based on the algorithm, for example, if we target those who have a CT infection probability of 0.082 or greater for testing, we could capture 83% of infected individuals while reducing the number of tests by more than a half. Similarly, with appropriately chosen cutoff points, desired levels of sensitivity could be achieved with greatly reduced number of testing for GC and TV infections. Therefore, the proposed targeted screening algorithm had an excellent performance in attaining high sensitivity as well as reducing testing cost.

7. Discussion

In this research, we propose a model-based algorithm to identify individuals who are at increased STI risk for targeted screening. The algorithm takes into account the concurrent and nonlinear influences of demographic and behavioral risk factors as well as clinical factors to estimate the probability of organism-specific infection acquisition. With this algorithm, screening decisions can be made based on estimated STI probabilities, instead of the simple presence or absence of individual risk indicators. For all practical purposes, one typically expects a good targeted screening program to have a decent level of sensitivity to ensure infected individuals receive STI testing. In reality, desired levels of sensitivity are usually achieved at the expense of reduced specificity. In other words, we favor a sensitive screening program that captures most infected individuals, at the price of testing individuals who do not have STI. Lower specificity tends not to be an issue as the ensuing biological tests will confirm the absence of infection, at an added price. For this reason, a risk measure of STI on the continuum of a probability range provides the necessary flexibility to balance the levels of sensitivity and screening cost. Our research demonstrates that it is possible to drastically reduce the number of tests while maintaining an excellent level of sensitivity. Therefore, the proposed model-based algorithm can potentially facilitate targeted STI screening and improve screening efficiency.

It is important to note that the proposed model does not require extensive amount of behavioral information. From a modeling perspective, sexual behavioral information, especially that concerns the characteristics of the sexual partners, can be useful predictors for infection. If available, those factors can be easily incorporated in the proposed modeling structure. Such expansion of the model is likely to further increase its predictive performance. This said, we want to minimize the burden of clinical data collection by reducing the number and intrusiveness of the behavioral questions. With these considerations in mind, we feel that an excellent predictive performance with limited information input is an important strength of the currently presented model.

Importantly, the model can also be used to address STI-related epidemiological questions. For example, with the proposed model we have been able to express CT, GC and TV infection risks as functions of age and the number of sexual partners in a comparative manner. Previous studies have examined the age trends of these common STIs [46, 47], but few studies have directly quantified age and organism-specific STI risks in longitudinal cohorts, possibly due to the lack of appropriate analytical tools. Our research has confirmed the differential timing of the peak risks for CT, GC and TV, with the respective peak ages at 14 – 16, 18 – 19, and 24 – 25 years. Furthermore, the waning partner effect on CT over age once again raises an important question about the underlying causes of the early emergence of CT infections, in comparison to the relatively late surge of TV infections [15, 48]. While the prevalence of these STIs in the partner population may in part explain the organism-specific timing of infection acquisition, this does not exclude the possibility of additional contributing factors, for example, cervico-vaginal tissue immaturity, cervical ectopy, and immunological näıvetẽ in younger women [49]. The latter explanation has become particularly attractive, considering the fact that we observed clearly different partner effects between younger and older participants. Clinically, these results can help better define the

risk profiles for those common STIs in young women and thus improving the efficiency of STI screening.

From a methodological standpoint, the proposed method provides a general framework for the analysis of multiple binary data in a longitudinal setting. The joint modeling framework has the flexibility to accommodate various types of dependency structure among multiple outcomes. The bivariate smoothing component allows for exploration of concurrent and potentially nonlinear effects of two independent variables. Along with the likelihood-based resampling procedure, the proposed modeling approach provides a practical tool to dissect the nonlinear influences and their interactions on multiple outcomes. As we have shown in the STI example, without such a tool, many of the important but more nuanced findings would not be made. Moreover, the method is generally applicable to a much wider class of biomedical applications where exploration of multiple biological influences is desired. The proposed modeling framework has been developed for binary outcomes, and it has the potential to be extended for other members in the exponential family, including multiple outcomes with different distributions. These extensions will further enhance the applicability of the proposed method. Finally, we emphasize that although the model accommodates cross-outcome dependency through random effects, the correlation parameters do not possess the usual interpretations of the correlation coefficients, nor can they be easily converted into the latter. More complex random effect structures could further complicate the parameters' interpretations. This said, if direct quantification of cross-outcome correlations is truly of interest, one may have to resort to an alternative model formulation, such as expressing the effect of one outcome as an odds ratio of the other. Notwithstanding this limitation, the present model provides an general approach for assessing nonlinear effects of risk factors on multiple binary outcomes.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgement

The authors are grateful to Dr. J Dennis Fortenberry, Department of Pediatrics, Indiana University School of Medicine, for his scientific advice on STI screening and for sharing his study data. The research is not possible without Dr. Fortenberry's valuable input.

Contract/grant sponsor: The research is funded by NIH grants RO1 HD042404, RO1 HD044387, and U19 AI 031494.

References

1. Cates W Jr. Estimates of the incidence and prevalence of sexually transmitted diseases in the United States. Sexually Transmitted Diseases. 1999; 26(4):S2–S7. [PubMed: 10227693]
2. Centers for Disease Control and Prevention. Sexually Transmitted Disease Surveillance 2010. U.S. Department of Health and Human Services; Atlanta: 2011.
3. Holmes, K.; Sparling, P.; Stamm, W.; Piot, P.; Wasserheit, J.; Corey, L.; Cohen, M. Sexually Transmitted Diseases. McGraw Hill Professional; 2007.
4. Centers for Disease Control and Prevention. Sexually Transmitted Disease Treatment Guidelines 2010. 2010. Morbidity and Mortality Weekly Report

5. US Preventive Services Task Force. Screening for chlamydial infection: U. S. Preventive Services Task Force recommendation statement. Annals of Internal Medicine. 2007; 147(2):128–134. [PubMed: 17576996]
6. Van Der Pol B, Ferrero DV, Buck-Barrington L, Hook E III, Lenderman C, Quinn T, Gaydos CA, Lovchik J, Schachter J, Moncada J, et al. Multicenter evaluation of the BDProbeTec ET system for detection of *Chlamydia trachomatis* and *Neisseria gonorrhoeae* in urine specimens, female endocervical swabs, and male urethral swabs. Journal of Clinical Microbiology. 2001; 39(3):1008–1016. [PubMed: 11230419]
7. Van Der Pol B, Quinn TC, Gaydos CA, Crotchfelt K, Schachter J, Moncada J, Martin DH, et al. Multicenter evaluation of the AMPLICOR and automated COBAS AMPLICOR CT/NG tests for detection of *Chlamydia trachomatis*. Journal of Clinical Microbiology. 2000; 38(3):1105–1112. D J. B T, C P. [PubMed: 10699004]
8. van Valkengoed IGM, Morré SA, van den Brule AJC, Meijer CJLM, Devillé W, Bouter LM, Boeke AJP. Low diagnostic accuracy of selective screening criteria for asymptomatic *Chlamydia trachomatis* infections in the general population. Sexually Transmitted Infections. 2000; 76(5):375–380. [PubMed: 11141855]
9. van Valkengoed IGM, Boeke AJP, Morré SA, van den Brule AJC, Meijer CJLM, Devillé W, Bouter LM. Disappointing performance of literature-derived selective screening criteria for asymptomatic *Chlamydia trachomatis* infection in an inner-city population. Sexually Transmitted Diseases. 2000; 27(9):504–507. [PubMed: 11034524]
10. Dicker LW, Mosure DJ, Berman SM, Levine WC, Regional Infertility Prevention Program. Gonorrhea prevalence and coinfection with Chlamydia in women in the United States, 2000. Sexually Transmitted Diseases. 2003; 30(5):572–576.
11. Kahn RH, Mosure DJ, Blank S, Kent CK, Chow JM, Boudov MR, Brock J, Tulloch S, Jail STD Prevalence Monitoring Project. *Chlamydia trachomatis* and *Neisseria gonorrhoeae* prevalence and coinfection in adolescents entering selected US juvenile detention centers, 1997-2002. Sexually Transmitted Diseases. 2005; 32(4):255–259. [PubMed: 15788927]
12. Fortenberry JD, Brizendine EJ, Katz BP, Wools KK, Blythe MJ, Orr DP. Subsequent sexually transmitted infections among adolescent women with genital infection due to *Chlamydia trachomatis*, *Neisseria gonorrhoeae*, or *Trichomonas vaginalis*. Sexually Transmitted Diseases. 1999; 26(1):26–32. [PubMed: 9918320]
13. Khan A, Fortenberry JD, Juliar BE, Tu W, Orr DP, Batteiger BE. The prevalence of Chlamydia, Gonorrhea, and Trichomonas in sexual partnerships: Implications for partner notification and treatment. Sexually Transmitted Diseases. 2005; 32(4):260–264. [PubMed: 15788928]
14. Weinstock H, Berman S, Cates W Jr. Sexually transmitted diseases among american youth: Incidence and prevalence estimates, 2000. Perspectives on Sexual and Reproductive Health. 2004; 36(1):6–10. [PubMed: 14982671]
15. Bernstein GR, Gaydos CA, Diener-West M, Howell MR, Zenilman JM, Quinn TC. Incident *Chlamydia trachomatis* infections among inner-city adolescent females. Journal of the American Medical Association. 1998; 280(6):521–526. [PubMed: 9707141]
16. Faber MT, Nielsen A, Nygård M, Sparén P, Tryggvadottir L, Hansen BT, Liaw KL, Kjaer SK. Genital Chlamydia, genital herpes, *Trichomonas vaginalis* and Gonorrhea prevalence, and risk factors among nearly 70,000 randomly selected women in 4 nordic countries. Sexually Transmitted Diseases. 2011; 38(8):727–734. [PubMed: 21844702]
17. Yu Z, Lin X, Tu W. Semiparametric frailty models for clustered failure time data. Biometrics. 2012; 68(2):429–436. [PubMed: 22070739]
18. Tu W, Ghosh P, Katz BP. A stochastic model for assessing *Chlamydia trachomatis* transmission risk using longitudinal observational data. Journal of the Royal Statistical Society. Series A (Statistics in Society). 2011; 174(4):975–989.
19. Rochon J. Analyzing bivariate repeated measures for discrete and continuous outcome variables. Biometrics. 1996; 52(2):740–750. [PubMed: 8672710]
20. Gray SM, Brookmeyer R. Estimating a treatment effect from multidimensional longitudinal data. Biometrics. 1998; 54(3):976–988. [PubMed: 9750246]

21. Carey V, Zeger SL, Diggle P. Modelling multivariate binary data with alternating logistic regression. Biometrika. 1993; 80(3):517–526.
22. O'Brien LM, Fitzmaurice GM. Analysis of longitudinal multiple-source binary data using generalized estimating equations. Journal of the Royal Statistical Society: Series C (Applied Statistics). 2004; 53(1):177–193.
23. Reinsel G. Multivariate repeated-measurement or growth curve models with multivariate random-effects covariance structure. Journal of the American Statistical Association. 1982; 77(377):190–195.
24. Shah A, Laird N, Schoenfeld D. A random-effects model for multiple characteristics with possibly missing data. Journal of the American Statistical Association. 1997; 92(438):775–779.
25. Coull BA, Staudenmayer J. Self-modeling regression for multivariate curve data. Statistica Sinica. 2004; 14:695–711.
26. Ruppert, D.; Wand, MP.; Carroll, RJ. Semiparametric Regression. Cambridge University Press; New York, NY: 2003.
27. Ghosh P, Tu W. Assessing sexual attitudes and behaviors of young women: A joint model with nonlinear time effects, time varying covariates, and dropouts. Journal of the American Statistical Association. 2009; 104(486):474–485.
28. Ghosh P, Hanson T. A semiparametric bayesian approach to multivariate longitudinal data. Australian & New Zealand Journal of Statistics. 2010; 52(3):275–288. [PubMed: 21731424]
29. Liu H, Tu W. A semiparametric regression model for paired longitudinal outcomes with application in childhood blood pressure development. The Annals of Applied Statistics. 2012; 6(4):1861–1882.
30. Tu W, Batteiger BE, Wiehe S, Ofner S, Van Der Pol B, Katz BP, Orr DP, Fortenberry JD. Time from first intercourse to first sexually transmitted infection diagnosis among adolescent women. Archives of Pediatrics & Adolescent Medicine. 2009; 163(12):1106–1111. [PubMed: 19996047]
31. Wood SN. Thin plate regression splines. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2003; 65(1):95–114.
32. Wood, SN. Generalized Additive Models: An Introduction with R. Chapman and Hall/CRC; Boca Raton, FL: 2006.
33. Barndorff-Nielsen, OE.; Cox, DR. Asymptotic Techniques for Use in Statistics. Chapman and Hall; London: 1989.
34. Bates DM. lme4: Mixed-effects modeling with R. 2010 http://lme4.r-forge.r-project.org/book/.
35. Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. Journal of the American Statistical Association. 1993; 88(421):9–25.
36. Wood SN. gamm4: Generalized additive mixed models using mgcv and lme4. 2011 URL http://CRAN.R-project.org/package=gamm4, r package version 0.1-5.
37. Schall N. Estimation in generalized linear models with random effects. Biometrika. 1991; 78(4): 719–727.
38. Rodriguez G, Goldman N. An assessment of estimation procedures for multilevel models with binary responses. Journal of the Royal Statistical Society. Series A (Statistics in Society). 1993; 158(1):73–89.
39. Goldstein H, Rasbash J. Improved approximations for multilevel models with binary responses. Journal of the Royal Statistical Society: Series A (Statistics in Society). 1996; 159(3):505–513.
40. Ng ESW, Carpenter JR, Goldstein H, Rasbash J. Estimation in generalised linear mixed models with binary outcomes by simulated maximum likelihood. Statistically Modelling. 2006; 6(1):23–42.
41. Zhang DW, Lin XH. Hypothesis testing in semiparametric additive mixed models. Biostatistics. 2003; 4(1):57–74. [PubMed: 12925330]
42. Crainiceanu CM, Ruppert D. Likelihood ratio tests in linear mixed models with one variance component. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2004; 66:165–185.
43. Härdle W, Huet S, Mammen E, Sperlich S. Bootstrap inference in semiparametric generalized additive models. Econometric Theory. 2004; 20:265–300.

44. Roca-Pardiñas J, Cadarso-Suárez C, Tahoces PG, Lado MJ. Assessing continuous bivariate effects among different groups through nonparametric regression models: An application to breast cancer detection. Computational Statistics & Data Analysis. 2008; 52:1958–1970.

45. Efron B. Bootstrap methods: Another look at the jackknife. The Annals of Statistics. 1979; 7(1):1–26.

46. Datta SD, Sternberg M, Johnson RE, Berman S, Papp JR, McQuillan G, Weinstock H. Gonorrhea and Chlamydia in the United States among persons 14 to 39 years of age, 1999 to 2002. Annals of Internal Medicine. 2007; 147(2):89–96. [PubMed: 17638719]

47. Sutton M, Sternburg M, Koumans EH, McQuillan G, Berman S, Markowitz L. The prevalence of *Trichomonas vaginalis* infection among reproductive-age women in the United States, 2001-2004. Clinical Infectious Diseases. 2007; 45(10):1319–1326. [PubMed: 17968828]

48. Miller WC, Swygard H, Hobbs MM, Ford CA, Handcock MS, Morris M, Schmitz JL, Cohen MS, Harris KM, Udry JR. The prevalence of Trichomonas in young adults in the United States. Sexually Transmitted Diseases. 2005; 32(10):593–598. [PubMed: 16205299]

49. Ethier, KA.; Orr, DP. Behavioral Interventions for Prevention and Control of STDs Among Adolescents. Springer; New York: 2007.

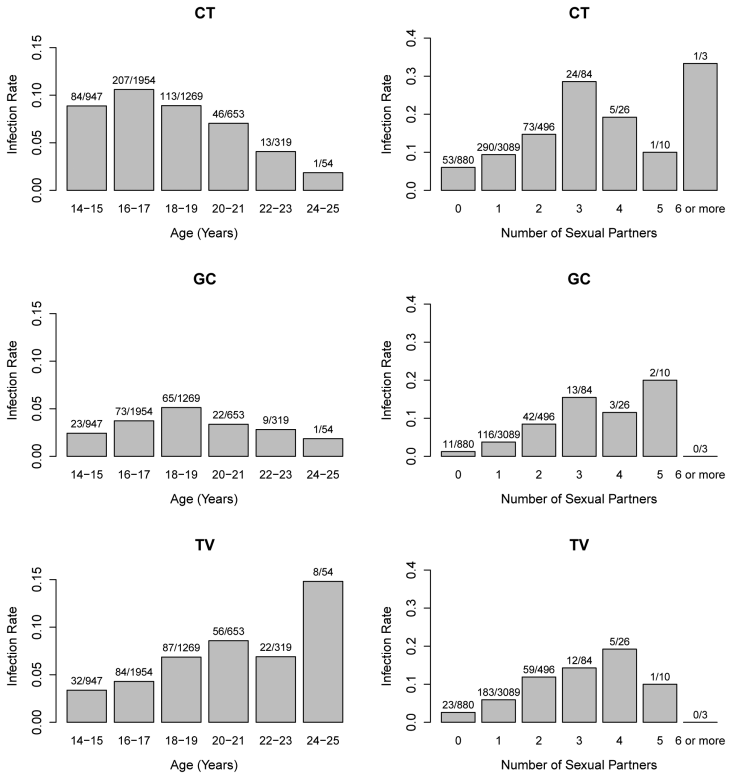

Figure 1.
CT, GC and TV infection rates by age and the number of sexual partners.

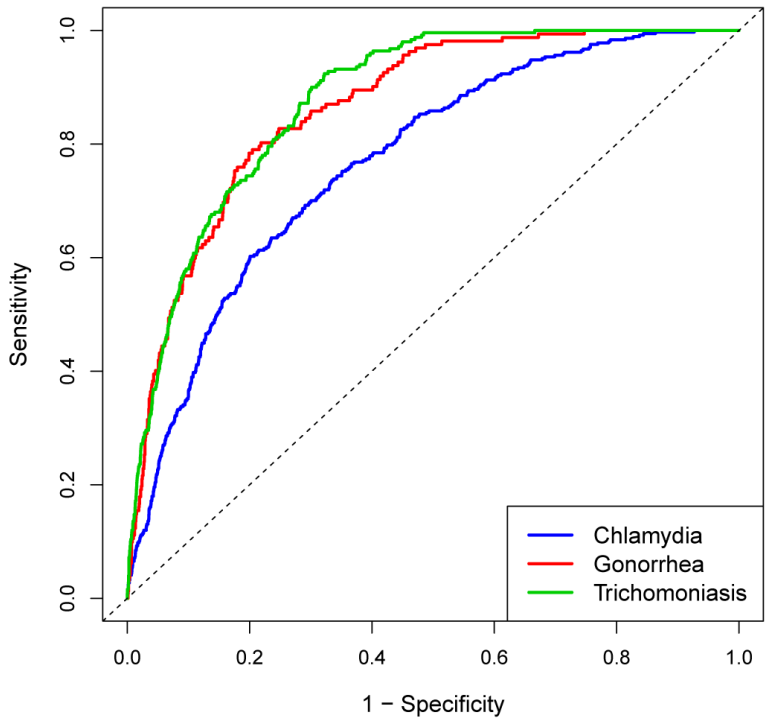

Figure 2.
ROC curves for CT, GC and TV with areas under the curves (AUC) of 0.80, 0.87 and 0.89 respectively.

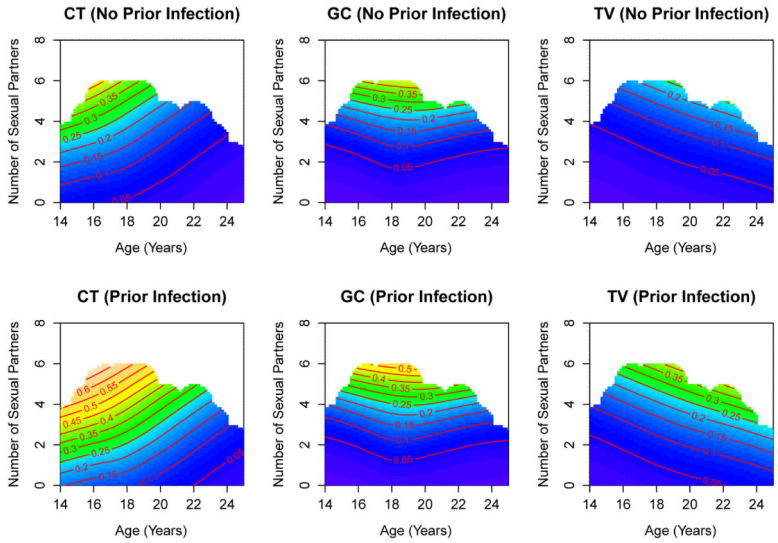

Figure 3.

Bivariate surfaces showing the joint effects of age and the number of sexual partners on CT, GC and TV infections with or without a prior infection.

Table 1

Parameter estimates and MSE of smooth functions averaged over 200 simulation runs. Average standard errors (in parentheses) and coverage probabilities of 95% CIs (in brackets) were calculated based on 200 bootstrap samples.

m	n	$\beta_0^1 = -2.5$	$\beta_0^2 = -3.5$	$\beta_1^1 = 1$	$\beta_1^2 = 0.7$	$\sigma_1 = 0.6$	$\sigma_2 = 1$	$\rho = 0.7$	MSE($\hat{f_1}$)	MSE($\hat{f_2}$)
200	10	−2.498 (0.113) [95.0%]	−3.569 (0.233) [91.0%]	1.026 (0.184) [91.5%]	0.641 (0.308) [93.5%]	0.562 (0.141) [97.0%]	1.054 (0.209) [92.5%]	0.693 (0.251) [94.5%]	0.0260	0.0282
200	20	−2.505 (0.082) [93.0%]	−3.529 (0.149) [88.0%]	1.011 (0.134) [94.0%]	0.674 (0.223) [94.5%]	0.572 (0.086) [91.0%]	1.009 (0.129) [90.5%]	0.699 (0.158) [93.0%]	0.0179	0.0160
400	10	−2.509 (0.080) [91.0%]	−3.584 (0.156) [87.0%]	1.001 (0.132) [95.0%]	0.663 (0.211) [93.5%]	0.587 (0.100) [91.0%]	1.076 (0.142) [86.5%]	0.649 (0.180) [93.0%]	0.0182	0.0168
400	20	−2.499 (0.058) [94.0%]	−3.515 (0.102) [90.0%]	0.999 (0.095) [94.0%]	0.701 (0.155) [94.0%]	0.584 (0.061) [93.5%]	0.991 (0.091) [94.5%]	0.681 (0.116) [91.5%]	0.0116	0.0083

Table 2

Parameter estimates for model (8) with bootstrap standard errors and 95% CIs.

Parameter	Estimate	Std. Error	95% CI
β_0^{ct}	−2.59	0.13	(−2.83,−2.38)
β_0^{gc}	−3.95	0.26	(−4.52,−3.60)
β_0^{tv}	−3.55	0.19	(−3.97,−3.19)
β_2^{gc}	0.07	0.12	(−0.20,0.31)
β_2^{gc}	0.42	0.23	(−0.10,0.87)
β_2^{tv}	0.17	0.17	(−0.14,0.49)
σ_1	0.60	0.12	(0.45,0.89)
σ_2	1.02	0.26	(0.53,1.84)
σ_3	1.17	0.20	(0.82,1.65)
ρ_{12}	0.75	0.16	(0.45,1.00)
ρ_{13}	0.45	0.18	(0.14,0.78)
ρ_{23}	0.43	0.18	(0.13,0.79)

Table 3

Sensitivity and specificity of the proposed model under different cutoff points of infection probability for CT, GC and TV, and percentages of follow-up visits that meet the cutoff points.

Organism	Cutoff Point	Sensitivity	Specificity	Percentage of visits (%)
CT	0.068	0.90	0.42	61
	0.082	0.83	0.55	49
	0.096	0.74	0.67	38
	0.126	0.60	0.80	24
GC	0.029	0.93	0.58	44
	0.051	0.80	0.78	25
	0.062	0.70	0.84	18
	0.074	0.60	0.89	13
TV	0.048	0.92	0.68	36
	0.066	0.80	0.77	27
	0.090	0.70	0.84	20
	0.118	0.60	0.89	14